Proof of Concept: Concept-based Biomedical Information Retrieval

Dolf Trieschnigg
University of Twente, Enschede, The Netherlands
trieschn@ewi.utwente.nl

Abstract

In this thesis we investigate the possibility to integrate domain-specific knowledge into biomedical information retrieval (IR). Recent decades have shown a fast growing interest in biomedical research, reflected by an exponential growth in scientific literature. An important problem for biomedical IR is dealing with the complex and inconsistent terminology encountered in biomedical publications. Dealing with the terminology problem requires domain knowledge stored in terminological resources: controlled indexing vocabularies and thesauri. The integration of this knowledge is, however, far from trivial.

The first research theme investigates heuristics for obtaining word-based representations from biomedical text for robust retrieval. We investigated the effect of choices in document preprocessing heuristics on retrieval effectiveness. Document preprocessing heuristics such as stop word removal, stemming, and breakpoint identification and normalization were shown to strongly affect retrieval performance. An effective combination of heuristics was identified to obtain a word-based representation from text for the remainder of this thesis.

The second research theme deals with concept-based retrieval. We compared a word-based to a concept-based representation and determined to what extent a manual concept-based representation can be automatically obtained from text. Retrieval based on only concepts was demonstrated to be significantly less effective than word-based retrieval. This deteriorated performance could be explained by errors in the classification process, limitations of the concept vocabularies and limited exhaustiveness of the concept-based document representations. Retrieval based on a combination of word-based and automatically obtained concept-based query representations did significantly improve word-only retrieval.

In the third and last research theme we propose a cross-lingual framework for monolingual biomedical IR. In this framework, the integration of a concept-based representation is viewed as a cross-lingual matching problem involving a word-based and concept-based representation language. This framework gives us the opportunity to adopt a large set of established cross-lingual information retrieval methods and techniques for this domain. Experiments with basic term-to-term translation models demonstrate that this approach can significantly improve word-based retrieval.

Directions for future work are using these concepts for communication between user and retrieval system, extending upon the translation models and extending CLIR-enhanced concept-based retrieval outside the biomedical domain.

Available online from `http://purl.utwente.nl/publications/72481`